Comparison of Spectra in Unsequenced Species

Freddy Cliquet, Guillaume Fertin, Irena Rusu, Dominique Tessier

Comparison of Spectra in Unsequenced Species

Freddy Cliquet[1,2], Guillaume Fertin[1], Irena Rusu[1],Dominique Tessier[2]

[1] LINA, UMR CNRS 6241 Université de Nantes,
2 rue de la Houssinière, 44322, Nantes, Cedex 03, France
{freddy.cliquet,guillaume.fertin,irena.rusu}@univ-nantes.fr
[2] UR1268 BIA, INRA, Rue de la Géraudière, BP 71627, 44316 Nantes, France
dominique.tessier@nantes.inra.fr

Abstract. We introduce a new algorithm for the mass spectrometric identification of proteins. Experimental spectra obtained by tandem MS/MS are directly compared to theoretical spectra generated from proteins of evolutionarily closely related organisms. This work is motivated by the need of a method that allows the identification of proteins of unsequenced species against a database containing proteins of related organisms. The idea is that matching spectra of unknown peptides to very similar MS/MS spectra generated from this database of annotated proteins can lead to annotate unknown proteins. This process is similar to ortholog annotation in protein sequence databases. The difficulty with such an approach is that two similar peptides, even with just one modification (i.e. insertion, deletion or substitution of one or several amino acid(s)) between them, usually generate very dissimilar spectra. In this paper, we present a new dynamic programming based algorithm: PacketSpectralAlignment. Our algorithm is tolerant to modifications and fully exploits two important properties that are usually not considered: the notion of inner symmetry, a relation linking pairs of spectrum peaks, and the notion of packet inside each spectrum to keep related peaks together. Our algorithm, PacketSpectralAlignment is then compared to SpectralAlignment [1] on a dataset of simulated spectra. Our tests show that PacketSpectralAlignment behaves better, in terms of results and execution time.

1 Introduction

In proteomics, tandem mass spectrometry (MS/MS) is a general method used to identify proteins. At first, during the MS/MS process, the peptides issued from the digestion of an unknown protein by an enzyme are ionized so that their mass may be measured. Then, the mass spectrometer isolates each peptide and fragments it into smaller ions, before measuring the corresponding masses. This process provides for each peptide an **experimental spectrum** in the form of a series of peaks, each peak corresponding to a mass that has been measured. From these experimental spectra, we aim at retrieving the corresponding peptide sequences. Then, by combining the peptide sequences from different spectra, our goal is to relate the unknown protein to one of the proteins stored in a protein

database such as SwissProt. Given an experimental spectrum, there are two main possibilities to obtain the corresponding peptide sequence: *de novo* sequencing or *spectra comparison*.

***De novo*:** In *de novo* sequencing, the spectrum is directly interpreted and a corresponding amino acid sequence of the peptide is inferred without using any other data than the spectrum itself. Currently, the analysis of unsequenced species is mainly done by *de novo* interpretation [2, 3]. The main reason is the speed of each interpretation and the fact that the peptide sequence does not need to be part of a protein database. An important drawback is that this method requires high quality spectra and still leads to a lot of interpretation errors [4, 5].

Spectra Comparison: In spectra comparison, an experimental spectrum is compared with theoretically predicted spectra. The theoretical spectra are inferred from the different peptides generated by an *in-silico* digestion of all the proteins contained in a database. A score function is used to evaluate each possible comparison, and the results are ordered according to this score. The theoretical spectrum with the best score is associated to the corresponding experimental spectrum. There are a number of existing softwares that match uninterpreted MS/MS experimental spectra to theoretical spectra, such as SEQUEST [6] or MASCOT [7]. They are based on the Shared Peaks Count (SPC) algorithm, an algorithm that simply counts the number of peaks in common between the two spectra, but presents zero tolerance to **modifications** (i.e. insertion, deletion or substitution of one or several amino acid(s)). This is due to the fact that the slightest modification in a peptide sequence highly changes the contents of the corresponding spectrum, and thus can lead to false identifications. Two approaches have already been explored to take modified peptides into account. The first one consists in extending the database by applying all the possible modifications to each peptide of the base. However, this solution, leading to an exponential number of possibilities, is of course too time consuming [8]. The other one, SpectralAlignment [1, 9, 10], is a dynamic programming algorithm that has been designed to identify peptides even in presence of modifications. This method works rather well for one or two modifications, but for a larger number of modifications, SpectralAlignment is not really sustainable [9].

Spectra comparison has an essential advantage, namely the precision in the comparison, allowing information to be drawn even from spectra which used to be unexploitable with a *de novo* approach. But spectra comparison has a major downfall: the running time, that is highly dependent on the protein database size used to infer theoretical spectra.

Given that the proteins we are trying to identify come from unsequenced species, the idea is to find similar proteins on phylogenetically related organisms. This is why our method will have to allow the correspondence of a spectrum with a slightly different peptide. Yet, we need results with few errors and for most spectra. Although *de novo* is specially designed to treat unsequenced species, it still leads to lots of misinterpreted or uninterpreted spectra. That is why the development of a new spectra comparison method tolerant to modifications appears interesting for us.

2 Definitions and Notations

An MS/MS spectrum is obtained by selection, isolation and fragmentation of peptides within the mass spectrometer. Each peak results from the dissociation of a given peptide into two fragmented ions: an N-terminal one, and a C-terminal one. The **N-terminal** ions (that are called a, b, c, see for instance Figure 1) represent the left part of the peptide sequence, and the **C-terminal** ions (that are called x, y, z) represent the right part of the peptide sequence. The position where the fragmentation appears is located around the junction of two amino acids, as described in Figure 1. The b and y ions mark the exact junction point. In the rest of this paper, we will say that two different ions are **dependent** if they are issued from the fragmentation between the same two successive amino acids inside a peptide. For instance, in Figure 1, for a given $i \in [1;3]$, a_i, b_i, c_i, x_i, y_i and z_i are mutually dependent ions. In a spectrum, a peak corresponding to a C-terminal (resp. N-terminal) ion is called C-terminal (resp. N-terminal). Moreover, peaks corresponding to dependent ions are called dependent peaks.

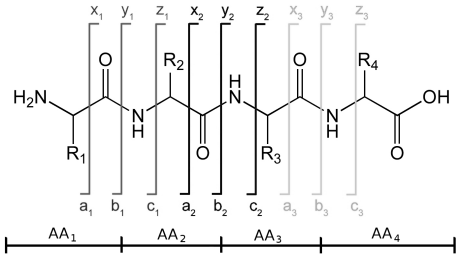

Fig. 1. This figure shows the fragmentation points inside a peptide containing four amino acids (AA_i with $i \in [1;4]$). R_i ($i \in [1;4]$) are chemical compounds that determine the corresponding amino acid. In this example, there are three sets of dependent ions.

We can notice that a symmetry exists between N-terminal and C-terminal peaks for a given fragmentation. In Figure 2 (a), the N-terminal peak located at position $m(GL)$ and the C-terminal peak located at position $m(MPRG)$ are linked by the relation $m(MPRG) = m(GLMPRG) - m(GL) - 20$ (-20 is due to the fact that the peptide is not symmetric at its extremities, see Figure 1, and to the ionization). This is a critical notion that is valid for any fragmentation and is seldom used. We call this relation **inner symmetry**.
Any spectrum produced by a mass spectrometer is called an **experimental spectrum** (S_e), and any spectrum predicted *in-silico* from a peptide is called a **theoretical spectrum** (S_t). In the following figures, spectra will always display all of the nine most frequent peaks coming from the fragmentation [11, 12].

Considering that the masses where the peaks appear can be seen as integers, we can represent a spectrum by a vector of booleans. This vector contains, for

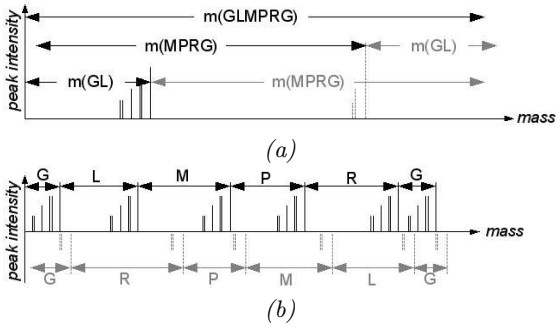

Fig. 2. (a) This piece of a spectrum shows the peaks created when the peptide GLMPRG (of mass $m(GLMPRG)$) is broken into peptides GL and MPRG. **(b)** is the spectrum of the peptide GLMPRG (for a better visualization, the N-terminal peaks are above the axis, the C-terminal ones are below the axis).

every mass, a boolean attesting the presence of a peak ('true') or its absence ('false'). Vector V_t represents the spectrum S_t and V_e represents S_e. Then, as in the case of sequences, we can align the elements of V_t and V_e two by two while allowing the insertion of gaps. It is only necessary to ensure that both vectors have the same length and that a gap is not aligned with another gap. In this representation, a **shift** corresponds to the insertion of a gap in either V_t or V_e, which itself corresponds to a peptide modification.
A **score** is used to evaluate an alignment, which represents the similarity between both spectra, in the sense that a higher score means a higher similarity. For instance, the number of common peaks in a given alignment is a possible score. When transposed to V_t and V_e, this corresponds to the number of pairs of booleans in which both values are 'true' at the same position. In this context, the alignment having the highest score will be considered as the best alignment. Our goal is, given an experimental spectrum, to compare it to all the theoretical spectra from the database. For each comparison, we want to find the best alignment.

3 Our Method

Because we want to align a spectrum S_e originated from unsequenced species with a spectrum S_t generated from phylogenetically related organisms, our method must be able to take modifications into account. Here, modification means insertion, deletion or substitution of one or several amino acids, but we will see that our algorithm, by nature, is able to handle other types of modifications such as post translational modifications.
Dynamic programming is an appropriate algorithmic method to find the best alignment between two spectra, even in presence of modifications. It has been used in this context by methods such as SpectralAlignement (SA) [1].

For our method to be tolerant to substitutions, we must be careful about the way we shift peaks when we want to improve our alignment, because a substitution, by changing the mass of an amino acid, not only changes one peak, but also some other peaks inside the whole spectrum. To take into account a substitution, a naïve algorithm could simply shift all peaks positioned after the location of the substituted amino acid. However, this could drastically change the positions of the peaks in the whole spectrum, as shown in Figure 3. It can be seen that such a modification destroys the link between N-terminal and C-terminal peaks, causing the inner symmetry to be broken. Moreover, the loss of information due to this rupture in the symmetry grows with the number of modifications.
In a previous work [1], Pevzner et al. considered this notion of symmetry by proposing, for each peak of each spectrum, to add its *symmetric twin* peak. However, they noticed two major drawbacks to this:

- The addition of noise (the symmetric twin of a noise peak is another noise peak).
- During the alignment, if a peak P (resp. a symmetric twin peak TP) is aligned, then TP (resp. P) must not be used in the alignment. Deciding, for each peak of the spectrum, which one of these peaks (P or TP) must be aligned, is an NP-complete problem [11].

But it is important to note that the two spectra S_e and S_t do not present the same properties. By construction, in S_t, we can identify each peak as an N-terminal or a C-terminal peak. This extra information is fully exploited in our algorithm and eliminates both previously raised drawbacks.

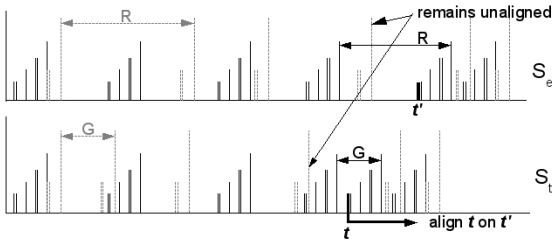

Fig. 3. In this figure, two spectra are represented. C-terminal peaks are dashed gray, N-terminal peaks are black. Spectrum S_e represents the peptide GLMPRG, spectrum S_t represents the peptide GLMPGG. This figure shows how one substitution in a peptide can highly change the contents of a spectrum (as it is, there are only 27 peaks aligned between S_e and S_t out of 54), and how shifting peaks from position t to position t' is not sufficient to correctly align S_e with S_t (the black arrow shows the shift, resulting in an alignment of 45 peaks between S_e and S_t out of 54). Although the shift improves the alignment, some of the peaks remain unaligned after the shift has been applied.

3.1 Symmetry

Theoretical Spectrum: We build the spectrum *in-silico*, so the location of all the different peaks is known. Considering this, we can easily remove all the C-terminal peaks and replace them by their respective symmetric peaks. Let m_i be the mass of the i-th C-terminal peak; its mass will be replaced by $M_{peptide} - m_i$ where $M_{peptide}$ is the mass of the peptide represented in S_t. The theoretical spectrum, after the application of the symmetry, is called **theoretical symmetric spectrum** (SS_t). Figure 4 (above) shows the spectrum S_t corresponding to the peptide *GLMPRG*. Figure 4 (below) shows the spectrum SS_t, that is the spectrum S_t to which the symmetry has been applied. Note that in SS_t, the distance between the N-terminal and C-terminal peaks from the same fragmentation is a constant (i.e. in spectrum SS_t from Figure 4, peaks t and t' (resp. u and u') are dependent and the distance between t and t' is the same than between u and u'), thus a shift of all the peaks positioned after a modification will still respect the *inner symmetry*. We also point out that our construction of SS_t is different that the one proposed by Pevzner et al. in [1] on two points: (1) we do not apply symmetry on the N-terminal peaks and (2) when applying symmetry on C-terminal peaks, we remove the original peaks.

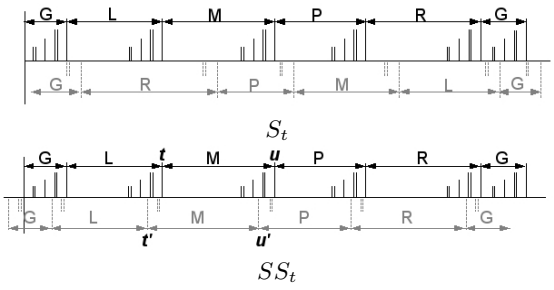

Fig. 4. S_t *(above)* represents the spectrum before any symmetry is applied. SS_t *(below)* represents the spectrum after the symmetry has been applied. In S_t the dashed peaks are the N-terminal peaks. In SS_t the dashed peaks are the N-terminal peaks to which symmetry has been applied.

Experimental Spectrum: As we do not know the ion type represented by each peak (in fact, a peak can represent the superposition of different ions of the same mass), we create a new symmetric peak for each existing peak of the spectrum. These peaks are created at position $M_{peptide} - m_i$, where $M_{peptide}$ represents the mass of the peptide measured by the mass spectrometer and m_i the mass of the i-th peak of S_e. The experimental spectrum, after the application of the symmetry is called **experimental symmetric spectrum** (SS_e). Figure 5 (above) shows the spectrum S_e corresponding to the peptide *GLMPGG*. Figure 5 (below) shows the spectrum SS_e, that is the spectrum S_e to which the symmetry has been applied.

During the alignment, we need to forbid the alignment of some pairs of peaks. For instance, N-terminal peaks from SS_t should be aligned with the original peaks from SS_e. This is important to guarantee that the solution is feasible. To do this, it is sufficient to keep track, for each peak, if it is a original peak or not.

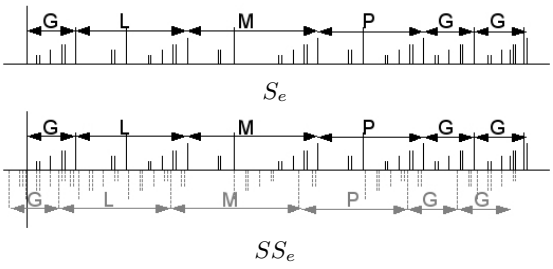

Fig. 5. S_e *(above)* represents the spectrum before any symmetry is applied. Then SS_e *(below)* represents the spectrum after the symmetry has been applied. In SS_e, the dashed peaks are S_e peaks to which symmetry has been applied.

3.2 Packet

The random fragmentation along the peptide backbone creates a number of different types of peaks. The location where the fragmentation occurs implies various peaks, a, b, c and x, y, or z (see Figure 1). Additionally, peaks corresponding to neutral loss (water, ammonia) are also frequently observed. For the construction of SS_t we choose to keep, for each type of fragmentation, the nine most frequent peaks observed in experimental spectra when using a *Quadrupole Time-of-Flight* mass spectrometer [11, 12] (see also Figure 6). After the application of symmetry, the notion of inner symmetry does not depend anymore of the peptide mass, thus these nine peaks may be clustered in a single **packet**. A packet represents all the fragmentations occurring between two consecutive amino acids of the peptide, thus SS_t can now be represented by a group of packets. An example of packet is shown in Figure 6. Point R_p marks the **reference point** of a packet p, and will be referred to when we will talk about a *packet position*. Introducing this notion of indivisible packet into the comparison between experimental and theoretical spectra allow us to forbid any translation that pulls apart dependent peaks: indeed, a shift can only exist between two packets.
In addition, to align SS_t with SS_e, for each packet p of SS_t, R_p is positionned on a mass m of SS_e. This gives an alignment of score s between p and SS_e. If s goes past a threshold, then the mass m is considered as one of the **possible masses**. In the rest of this paper, this score will be the number of aligned peaks. Increasing the threshold T will speed up the alignment process because the number of possible masses will decrease (see Table 1).

Another constraint which forbids the overlapping of two packets is added; it represents the fact that an amino acid has a minimum mass. That way, we do not allow the algorithm to make small and unrealistic shifts just to slightly improve its score (something which happens very often with SA).
Note that it is possible to modify the contents of a packet in order to adapt this notion for other types of mass spectrometers.

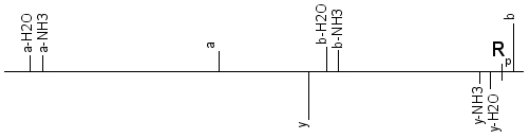

Fig. 6. The nine peaks representing the nine most current peaks that occur in MS/MS spectra (a, b, y and some of their variant peaks that are due to water or ammonia loss) result after the symmetry has been applied, to this packet. This packet is particularly suited for *Quadrupole Time-of-Flight* mass spectrometer.

3.3 PacketSpectralAlignment Algorithm

Our PacketSpectralAlignment (PSA) method needs three parameters: (i) SS_e as described in Section 3.1, (ii) SS_t for which the peaks are clustered into packets as described in Section 3.2 and (iii) K, the maximum number of allowed shifts. Our PSA algorithm (Algorithm 1) uses two matrices M and D. The value $M[p][m][k]$ represents the best score obtained so far for the alignment of the peaks of the first packets of SS_t, up to packet number p, with all the first peaks from SS_e, up to the alignment of R_p, with mass m, and with at most k shifts. The value $D[p][m][k]$ represents the best score of an alignment that must have R_p aligned with mass m in SS_e and containing at most k shifts. PSA will compute, for each possible number k of shifts, each possible mass m in SS_e and for each packet p of SS_t, the values $M[p][m][k]$ and $D[p][m][k]$.
The variable *best* contains the score of the best alignment met on the D matrix that could be extended with the current position (p, m, k) without more shifts (see Figure 7 for an illustration of how one of these alignments is found).
The *Score* function will return the score resulting from the alignment of the peaks of the p-th packet of SS_t with the peaks of SS_e, when R_p is positioned at the mass m of SS_e.
The D matrix is updated by choosing the best possibility between the two following cases:

a. we used the last value met on the diagonal, meaning no shift is needed, or

b. we must apply a shift and take the last score met on the diagonal in the $(k-1)$-th dimension of the matrix.

Then the M matrix is updated by taking the best alignment found until this

point.
As an illustration, applying our algorithm PSA on the theoretical spectrum SS_t of Figure 4 and the experimental spectrum SS_e of Figure 5 gives a "perfect" alignment of the peaks (i.e. that is, 54 peaks out of 54 that are aligned) with one shift (corresponding to the substitution of the fifth amino acid R by G in SS_t).

Algorithm 1 PSA(ExperimentalSpectrum SS_e, TheoreticalSpectrum SS_t, Integer K)

Ensure: The best alignment between SS_e and SS_t with a max of K shifts

for $k = 0$ to K **do**
  for all Possible masses m from SS_e **do**
    for all Packets p from SS_t **do**
      $best = \max\{D[p'][m - (R_p - R_{p'})][k] \mid p' < p\}$ /*(see Figure 7)*/
      $s = \text{Score}(p,m)$
      $D[p][m][k] = \max(best+s, M[p-1][m-PacketSize][k-1]+s)$ /*$PacketSize$ is the constant representing the size of a packet (i.e. the distance between the first and the last peak of a packet)*/
      $M[p][m][k] = \max(D[p][m][k], M[p-1][m][k], M[p][m-1][k])$
    end for
  end for
end for
return $M[NbPacket][MAX][K]$ /*$NbPacket$ is the number of packets composing SS_t and MAX is the highest mass represented by a peak inside SS_e*/

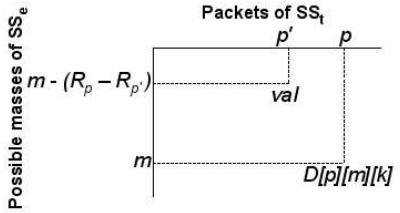

Fig. 7. This figure shows how to find a value $val = D[p'][m - (R_p - R_{p'})][k]$, with $p' < p$.

4 Results

We compare our algorithm to SA on a set of simulated data. We generate a dataset of 1000 random peptides of random size in $[10, 25]$ in order to constitute a database that will be used to create the theoretical symmetric spectra. Each

peptide in the database is then modified in 5 different versions by applying 0 to 4 random substitutions of amino acids. These modified peptides are used to create 5 sets of artificial experimental symmetric spectra (one for each different number of modifications). These spectra are constituted using the nine most frequents peaks that are created considering the probability of apparition observed by Frank et al. [12]. Noise has been introduced in each spectra, adding 50% more peaks at random masses. All tests have been made using 1 *dalton* precision. For each SS_e, we call the **target peptide** of SS_e (denoted $TP(SS_e)$) the original peptide sequence from the dataset that has been modified to obtain the spectrum.
Each SS_e is compared with each SS_t. The score (here, the number of common peaks in the best alignment) resulting from each comparison is memorized and used to order the set of peptides for each SS_e. In the ideal case, for an experimental spectrum SS_e, $TP(SS_e)$ should obtain the highest score and thus should have rank one. Thus, by looking at the rank of the target, we can evaluate the capacity for PSA to take modifications into account. That is why we use the average rank of the target peptide, as well as the proportion of target peptides having rank one, as indicators to evaluate the two algorithms (as shown in Figure 8).

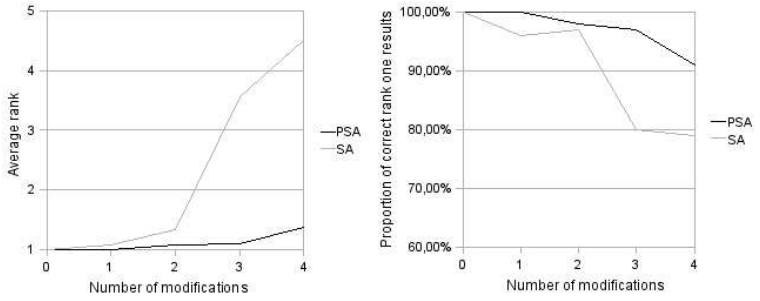

Fig. 8. Comparison of SA and PSA on our sets of 1000 random peptides

During the comparisons, the parameters used by SA and our algorithm PSA are the same. In particular, we use the same score function, which is the number of common peaks in both spectra. The number of shifts used by these two methods is set dynamically, depending on the size of the two compared spectra. We could have fixed this to a constant value, but allowing for instance $\frac{N}{2}$ shifts in an N amino acids long peptide does not make any sense, so we chose to allow k shifts for a peptide of mass M where $k = \lceil \frac{M}{600} \rceil + 1$. In the case of PSA, the threshold T used to determine the possible masses kept in symmetric experimental spectra is set to 2.
Our tests show that the two algorithms have a comparable behaviour for 0 to 2 shifts, with a slight advantage for our algorithm. However, for more than two

shifts, SpectralAlignment presents a fast deterioration of its results, while PacketSpectralAlignment still gives good results (see Figure 8). We also note that on these tests, for a threshold T of 2, our algorithm PSA is twice as fast as SA.

We have also evaluated the benefits supplied by the packets, and more particularly by the number of possible masses. As explained in Section 3.2, we do not test all masses in SS_e, but only those masses m inducing an alignment of at least T peaks when the reference point R_p of a packet p from SS_t is positionned at mass m. To evaluate this, we have computed the number of possible masses for different values of T on four different datasets. The first one is a set of 1000 simulated spectra of size $[10, 25]$ with 50% of noise peaks, generated the same way as described at the beginning of Section 4. On this dataset, a spectrum contains on average 150 peaks. Then we use three sets of 140 experimental maize spectra on which we apply different filters: (1) we keep all peaks (meaning an average of 275 peaks per spectrum), (2) we keep the 100 most intense peaks, and (3) we keep the 50 most intense peaks. Table 1 shows the evolution of the number of possible masses in function of the threshold T for each set of spectra. We can notice that the number of possible masses decreases considerably when T is increased.

Number of Possible Masses					
		Threshold T			
		1	2	3	4
Simulated spectra		485	134	39	14
Experimental Maize spectra	*no filtering*	689	312	141	61
	100 most intense peaks	540	180	57	17
	50 most intense peaks	346	79	18	4

Table 1. Evaluation of the number of possible masses on four sets of spectra depending on the threshold T.

5 Conclusion

We have developed PacketSpectralAlignment, a new dynamic programming based algorithm that fully exploits, for the first time, two properties that are inherent to MS/MS spectra. The first one consists in using the *inner symmetry* of spectra and the second one is the grouping of all dependent peaks into *packets*. Although our algorithm was at first motivated by the identification of proteins in unsequenced organisms, it does not set any constraints on the allowed shifts in the alignment. Thus, PSA is also able to handle the discovery of post translational modifications.
Our results are very positive, showing a serious increase in peptides identification in spite of modifications. The sensibility has been significantly increased,

while the execution time has been divided by more than two. More tests on experimental data will allow us to evaluate more precisely the benefits provided by our new algorithm. In the future, a better consideration of other points, such as spectra quality, will be added. Moreover, the score will be improved by taking into account other elements such as peaks intensity.

Acknowledgments. MS/MS experimental spectra were performed with the facilities of the platform Biopolymers, Interactions and Structural Biology, INRA Nantes. The authors thank Dr Hélène Rogniaux for fruitful discussions about MS/MS spectra interpretation. This research was supported by grant from the Region Pays de la Loire, France.

References

1. Pevzner, P.A., Dancík, V., Tang, C.L.: Mutation-tolerant protein identification by mass spectrometry. J Comput Biol **7**(6) (2000) 777–87
2. Habermann, B., Oegema, J., Sunyaev, S., Shevchenko, A.: The power and the limitations of cross-species protein identification by mass spectrometry-driven sequence similarity searches. Mol Cell Proteomics **3**(3) (2004) 238–49
3. Grossmann, J., Fischer, B., Baerenfaller, K., Owiti, J., Buhmann, J.M., Gruissem, W., Baginsky, S.: A workflow to increase the detection rate of proteins from unsequenced organisms in high-throughput proteomics experiments. Proteomics **7**(23) (2007) 4245–54
4. Pevtsov, S., Fedulova, I., Mirzaei, H., Buck, C., Zhang, X.: Performance evaluation of existing de novo sequencing algorithms. J Proteome Res **5**(11) (2006) 3018–28
5. Pitzer, E., Masselot, A., Colinge, J.: Assessing peptide de novo sequencing algorithms performance on large and diverse data sets. Proteomics **7**(17) (2007) 3051–4
6. Eng, J., McCormack, A., Yates, J.: An approach to correlate tandem mass spectral data of peptides with amino acid sequences in a protein database. J. Am. Soc. Mass Spectrom **5**(11) (1994) 976–989
7. Perkins, D.N., Pappin, D.J., Creasy, D.M., Cottrell, J.S.: Probability-based protein identification by searching sequence databases using mass spectrometry data. Electrophoresis **20**(18) (1999) 3551–67
8. Yates, J.R., Eng, J.K., McCormack, A.L., Schieltz, D.: Method to correlate tandem mass spectra of modified peptides to amino acid sequences in the protein database. Anal Chem **67**(8) (1995) 1426–36
9. Pevzner, P.A., Mulyukov, Z., Dancik, V., Tang, C.L.: Efficiency of database search for identification of mutated and modified proteins via mass spectrometry. Genome Res **11**(2) (2001) 290–9
10. Tsur, D., Tanner, S., Zandi, E., Bafna, V., Pevzner, P.A.: Identification of post-translational modifications by blind search of mass spectra. Nat Biotechnol **23**(12) (2005) 1562–7
11. Dancik, V., Addona, T., Clauser, K., Vath, J., Pevzner, P.: De novo peptide sequencing via tandem mass spectrometry. Journal of Computational Biology **6**(3-4) (1999) 327–342
12. Frank, A.M., Savitski, M.M., Nielsen, M.L., Zubarev, R.A., Pevzner, P.A.: De novo peptide sequencing and identification with precision mass spectrometry. J Proteome Res **6**(1) (2007) 114–23